科学大数据
——国家大数据战略的基石

郭华东

中国科学院遥感与数字地球研究所 北京 100094

摘要 作为人类的新型战略资源，大数据已成为知识经济时代的战略高地。其少量依赖因果关系、主要依靠数据相关性发现知识的新模式，使得其成为继经验、理论和计算模式之后的数据密集型科学范式的典型代表，带来了科研方法论的变革，正成为科学发现的新引擎。科学大数据作为大数据的重要分支，具有不可重复性、高度不确定性、高维性及计算分析高度复杂性的内部特征，以及在数据内容、数据体量、数据获取、数据分析等方面的外部特征，这给科学大数据的处理技术与方法提出了新的挑战。在以上分析基础上，文章提出了科学认知科学大数据，建设科学大数据基础设施，建立科学数据研究中心，以及构建科学大数据学术平台等建议。

关键词 大数据，科学大数据，数据驱动，数据密集型科学

DOI 10.16418/j.issn.1000-3045.2018.08.001

1 蓬勃发展的大数据

2013年7月17日，习近平总书记指出："浩瀚的数据海洋就如同工业社会的石油资源，蕴含着巨大生产力和商机。谁掌握了大数据技术，谁就掌握了发展的资源和主动权。"大数据已成为信息主权的一种表现形式，将是继边防、海防、空防之后大国博弈的另一个空间[1]。大数据正在改变人类生活和对世界的深层理解。

第二次工业革命的爆发，导致以文字为载体的数据量约每10年翻一番；从工业化时代进入信息化时代，数据量每3年翻一番。当前，新一轮信息技术革命与人类社会活动交汇融合，半结构化、非结构化数据的大量涌现，数据的产生已不受时间和空间的限制，引发了数据爆炸式增长，数据类型繁多且复杂，已经超越了传统数据管理系统和处理模式的能力范围[2]，人类正在开启大数据时代新航程。据国际数据公司（IDC）发布的2017年大数据白皮书预测，2025年全球大数据规模将增长至163 ZB，相当于2016年的10倍，大数据继续表现出更为强健的增长态势[3]。中国拥有的数据在国际上举足轻重，截至2012年，已占全球的13%，预计到2020年将产

资助项目：中国科学院战略性先导科技专项（A类）（XDA19030000）
修改稿收到日期：2018年8月13日

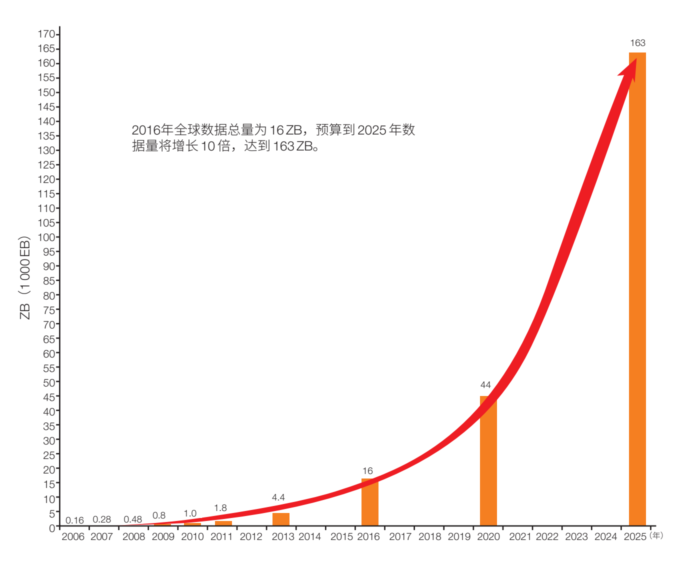


图 1　2016—2025年的全球数据量增长情况[3]

生全球20%的数据[4]。

从大数据搜索热度数据可清晰看出近年来全球对大数据的关注程度。国际上对大数据的关注度在2012年之前处于较低水平，2012—2015年对大数据的关注度飞速增长，2016年至今保持接近100的关注度。

国际上，从联合国到各国政府竞相重视大数据发展；在我国，大数据被列为国家战略后发展迅猛。全球大数据的发展方兴未艾，大数据已经开始显著地影响全球的生产、流通、分配和消费方式，它正在改变人类的生产方式、生活方式、经济运行机制和国家治理模式，它是知识驱动下经济时代的战略制高点，是国家和人类的新型战略资源。

2 科学大数据的认识

作为大数据的一个分支，科学大数据正在成为科学发现的新型驱动力，引起有关国家和科技界的高度重视。欧盟提出“科学是一项全球性事业，而科研数据是全球的资产”的理念[5]。美国的“从大数据到知识”计划、欧盟的“数据价值链战略计划”、英国的“科研数据之春”计划、澳大利亚的“大数据知识发现”项目、欧洲“地平线2020”计划的“数据驱动型创新”课题，均聚焦于从海量和复杂的数据中获取知识的能力，深入研究基于大数据价值链的创新机制，倡导大数据驱动的科学发现模式。大数据的影响已触及自然科学、社会科学、人文科学和工程科学的各个研究领域，不同领域的大数据研究中心陆续成立[6]。我国部署了一系列大数据科技项目，组建了不同研究方向的大数据实验室，中国科学院推出了“科学大数据工程”计划。

科学大数据具有数据密集型范式的特点，它具有数据的不可重复性、数据的高度不确定性、数据的高维特性、数据分析的高度计算复杂性等特征[7]。利用大量数据的相关性可取代因果关系和理论与模型，基于数据间的相关性能够获得新知识、新发现[8]。比如，早在1609年，第谷·布拉赫的助手约翰尼斯·开普勒从布拉赫对天体运动的系数观察记录中发现了行星运动定律，并发表了伟大的著作《新天文学》；又如，欧洲大型强子对撞机帮助物理学家检验关于不同粒子物理和高能物理理论的猜想，并且确定了希格斯玻色子的存在；再如，大数据使基因组学的科学发现成为可能；还如，时空大数据在全球环境研究变化中正发挥重大作用[9]。

越来越多的科学发现证明，大科学装置是人类认识自然世界的重要手段。对地观测卫星、大型望远镜、大型强子对撞机、高通量科学仪器、传感器网络等一系列大装置的成功运行，使得科学大数据与大装置和大科学间的关系越发密切。近年来，我国的大装置诸如500 m口径球面射电望远镜、系列空间科学卫星等的问世，为通过科学大数据认知大自然提供了强大的基础。为满足庞大且日益快速增长的科学大数据的应用需求，迫切需要建立一些能够共享数据、算法、模型的开放系统，以此实现对已有数据的科学分析和集成应用。一个典型的例子是，2017年10月，欧洲航天局“哨兵-5P”卫星发射后，每天获取近2 000万条空气污染物及气体的观测数据，其数据获取量是前期任务的10倍以上。按照目前的处理速度，一台计算机需要1 200年才能处理完300万景全球卫星影像。而基于云计算设施，可在45天内完成相同处理任务，足见重大基础设施的重要性[10]。

真正实现科学大数据的大价值尚面临着一系列技术挑战。在数据规模、数据增速、数据类型、数据质量、数据价值等方面给科学大数据处理技术与方法提出了新的科学技术问题和方向。

以上主要体现在5个方面：① 数据存储管理方面。科学大数据本身固有的特征亟待面向海量、非结构化或半结构化数据高效存储管理的数据库。② 数据分析方法方面。数据产生和数据分析过程的分离使得数据噪声增多，问题驱动的研究方式逐渐被数据驱动的研究方式所代替。③ 模型和算法方面。随着半结构化、非结构化数据比重的逐渐增多，针对该类数据的特征学习方法逐渐超越并取代传统的数据模型和算法。④ 计算体系结构方面。新型存储器件和计算器件不断涌现，使得通用处理器和单一体系结构的单机逐渐过渡为专用处理器、多核和分布式大规模异构集群。⑤ 计算和服务方面。以互联网为媒介的云计算模式和分布式高性能数据中心逐渐成为大数据处理的新型模式[2]。

中国科学院正在开展科学大数据研究的一些实践。如正在进行的中国科学院战略性先导科技专项（A类）“地球大数据科学工程”，地球大数据是一种典型的科学大数据，是具有空间属性的地球科学大数据。该专项力求突破超大规模跨域分布式资源技术瓶颈问题，有效推动地球大数据技术创新、聚合多时空数据管理与关联融合以及问题导向数据挖掘与分析，以达到只要有终端和互联网，任何人在任何地点都可以享受到地球大数据提供的多样服务，实现重大科学发现和一站式全方位宏观决策支持服务的目的[11]。

又如基于科学大数据的国际科学计划。我们于2016年发起的“数字丝路”（DBAR）国际计划，就是要实现大数据汇集、大数据服务、大数据分析和大数据呈现支撑，形成“一带一路”科学大数据平台。这个为期10年的科学计划，将为“一带一路”可持续发展、粮食安全、生态环境保护、气候变化监测、灾害风险应对，以及文化—自然遗产保护与发展等提供科学决策[12]。

再如基于科学大数据的研究项目。联合国设立了一项名为“全球脉动”的计划，其使命之一是用大数据应对气候挑战。2014年，在联合国气候变化峰会召开之际，来自46个国家的大数据应对气候变化项目参加了奖项竞争“挑战”，最终9个项目获得不同的奖励。我们的“对地观测大数据应对全球变化”研究项目入选其中，显示了空间对地观测大数据在气候变化研究中的作用和价值。

科学大数据正深刻改变传统的科研模式，正驱动现代科学研究的迅猛发展。科学大数据正在为科技创新带来大机遇。作为少量依赖因果关系，而主要依靠相关性发现新知识的新模式，科学大数据已成为继经验、理论和计算模式之后的数据密集型科学范式的典型代表[8]。

3 科学大数据的思考

随着数据积累和计算能力的提升，直接从大数据中获取知识已经成为可能。2013年9月，笔者及团队提出“科学大数据”概念，并于2014年1月以“科学大数据与数字地球”为题发表于《科学通报》。我们认为，科学大数据与互联网大数据、商业大数据等存在本质属性和特点上的区别，具有自己独特的科学内涵和特点[9]。

整体看来，科学大数据具有如下外部特征：从数据内容来讲，科学大数据一般表征自然客观对象和变化过程；从数据体量来讲，科学大数据在不同学科中存在较大的差异；从数据增长速率来讲，科学大数据依学科不同其数据增长速率也变化较大；从数据获取手段来讲，科学大数据一般来自观测和实验的记录以及后续加工；从数据分析手段来讲，科学大数据的知识发现一般需要借助科学原理模型。

通过归纳科学大数据的外部特征，其内部特征也变得相对清晰，主要概括为：数据内容的不可重复性。正如哲学家赫拉克利特的名言“人不能两次踏进同一条河流”，对于一般自然与物理的客观过程的观测具有一定的不可重复性。数据的高度不确定性。由于采用的直

接或非直接观测方式、采样手段和记录技术，往往引入系统观测误差及数据记录误差。数据的高维特性。由于观测对象和采样方法本身的时间、空间属性以及观测传感器的多通道特征，科学大数据往往具有时空连续性和谱段多维性，导致维数灾难。数据分析的高度计算复杂性。数据的高度不确定性、高维特性，以及与科学数据分析相伴随的原理模型的复杂性，导致了科学数据处理分析的计算复杂性。总之，科学大数据具有不同于一般大数据的特征，其内在机理及如何应用于知识发现需深入研究[7]。

2014 年 6 月，在我们的倡议和主持下，“国际科学计划大数据研讨会：挑战与机遇”在北京召开。该会议由国际科学和技术数据委员会（CODATA）主办，7 个国际组织共同主办。会议发表的声明强调科学研究要加强对大数据的理解，通过发展与大数据有关的研究、政策和框架来强化国际大数据科学合作，促进社会发展。尽管这在当时只是一个起点，但这份声明是人们关注大数据潜力迈出的实质性一步。声明要点包括：响应大数据对国际科学计划的重要性；开发大数据为社会服务的潜力；通过国际合作来增进对大数据的理解；通过全球研究基础设施促进大数据的普及；探索和应对大数据管理工作带来的挑战；鼓励大数据科学能力建设；促进政策制定，最大限度地利用大数据。

自那时起，我们主办或共同主办了一系列关于科学大数据的会议，其中包括“科学大数据前沿香山科学会议”“中国科学院学部空间地球大数据科学与技术前沿论坛”“自然科学与人文科学大数据前沿探索圆桌会议”“地球大数据香山科学会议”等。有关部门和单位相续组织召开了不同的与科学大数据有关的会议，进行深入研讨。

特别重要的是，在中国科学院的组织下，我们提出发展“科学大数据”的建议，上报后受到政府的重视。2015 年《国务院关于印发促进大数据发展行动纲要的通知》中把科学大数据作为纲要的一部分，提出“发展科学大数据：积极推动由国家公共财政支持的公益性科研活动获取和产生的科学数据逐步开放共享，构建科学大数据国家重大基础设施，实现对国家重要科技数据的权威汇集、长期保存、集成管理和全面共享。面向经济社会发展需求，发展科学大数据应用服务中心，支持解决经济社会发展和国家安全重大问题”[13]。

科学大数据是国家大数据战略的有机组成，这使得深入开展科学大数据的研究具备了良好的政策支撑和理论基础。科学大数据是国家大数据战略的基石，科技界和科学家肩负重大的使命——推进科学大数据的全面系统发展。

4 发展科学大数据的建议

全球范围内大数据蓬勃发展，我国正在实施国家大数据战略，科学大数据已成为大数据国家战略的重要组成部分。在习近平总书记对实施国家大数据战略提出更高要求的大背景下，国务院办公厅 2018 年 3 月又发布了《科学数据管理办法》。我们迎来了发展科学大数据的重要的历史机遇。为更好地推动科学大数据发展，有以下 4 点建议。

（1）科学认知大数据世界的科学大数据。大数据世界的科学大数据具有独到的特点，科学大数据提供了创新的科研方法论，科学大数据是驱动科学发现的新引擎，科学大数据是占领未来科学制高点的前沿领域，科学大数据为人类认识世界提供了全新的思维，科学大数据是孕育新型科学家的摇篮。目前，我国的计算机用户数全球第一，互联网用户数全球第一，移动互联网用户数全球第一，我国拥有的数据量未来几年有可能达到 20%，我国发表的大数据论文数目前国际排名第二。我国政府对大数据高度重视，我国的大数据在国际上有较高的话语权，为在此基础上开展的科学大数据研究走向国际前沿奠定了坚实基础。

（2）建设科学大数据国家重大基础设施。大装置产出大数据，大数据孕育大科学，大科学驱动大发现，

国家统一布局建设科学大数据国家重大基础设施十分重要。其意义包括保证科学大数据的获取与更新、权威汇集与高效处理，实现对重要科技数据的长期保存和集成管理。同时，科研活动过程中产生的海量科学数据需要通过网络给科学家们进行分析和处理，但海量数据的共享和传输过程，在当前的网络信息安全环境和条件下，导致科研数据传输的效率低下，影响科学发现的质量。进行科学大数据的收集、存储、维护、管理、分析和共享等核心技术需要重大基础设施的支撑。

（3）建立国家科学大数据研究中心。我国目前有数十个大科学装置、数百个国家重点实验室、大量的部门重点实验室，正在建设国家实验室。这些应是科学大数据首先“发力”的地方。建立科学大数据中心，服务于不同领域科研机构。可设立不同科学领域中心，如生命大数据中心、地球大数据中心、天文大数据中心等，开拓诸如生物信息学、地球信息学、天文信息学等相应的学科领域；也可设立不同区域科学大数据中心。考虑到中国科学院的国家定位，建议依托中国科学院建立国家科学大数据研究中心。同时，科学大数据能否顺利发展的关键之一是数据共享，应实施可持续发展的科学数据共享，包括重视科学数据出版这种新的数据集成与开放共享机制。

（4）发起科学大数据国际论坛与国际联盟。提高科学大数据在实践应用中的方法论、理论基础和技术研究，开展双边或多边的国际交流与合作是提高科学大数据研究水平的重要途径之一。国际科学论坛是保障以上实施的重要平台，有利于开展前沿理论的探讨，有利于加强与国际科技组织及国际科学计划的协作，以汇集更多领域、更多学科的专家力量，保持优良的国际科技合作环境。同时，应考虑建立国际科学大数据联盟。例如，面向“一带一路”倡议，构建大数据联盟。以科学大数据为抓手，让大数据成为“一带一路”建设的一个引擎，让大数据成为各国共建的和平使者，让大数据之光普照现在和未来。

致谢 梁栋同志为本文做了大量工作，特此感谢。

参考文献

1 Guo H D, Wang L Z, Chen F, et al. Scientific big data and digital earth. Science Bulletin, 2014, 59(35): 5066-5073.

2 郭华东, 陈润生, 徐志伟, 等. 自然科学与人文科学大数据——第六届中德前沿探索圆桌会议综述. 中国科学院院刊, 2016, 31(6): 707-716.

3 Reinsel D, Gantz J, Rydning J. Data age 2025: The evolution of data to life-critical don’t focus on big data. Framingham: IDC Analyze the Future, 2017.

4 Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. Framingham: IDC Analyze the Future, 2014.

5 GRDI2020. Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. [2018-08-16]. http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf.

6 李学龙, 龚海刚. 大数据系统综述. 中国科学：信息科学, 2015, 45(1): 1-44.

7 郭华东. 大数据、大科学、大发现——大数据与科学发现国际研讨会综述. 中国科学院院刊, 2014, 29(4): 500-506.

8 Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery. Washington DC: Microsoft Research, 2009.

9 郭华东, 王力哲, 陈方, 等. 科学大数据与数字地球. 科学通报, 2014, 59(12): 1047-1054.

10 Guo H D. Steps to the digital Silk Road. Nature, 2018, 554: 25-27.

11 Guo H D. Big Earth data: A new frontier in Earth and information sciences. Big Earth Data, 2017, 1(1-2): 4-20.

12 Guo H D, Liu J, Qiu Y B, et al. The Digital Belt and Road program in support of regional sustainability. International Journal of Digital Earth, 2018, 11(7): 657-669.

13 中华人民共和国国务院. 国务院关于印发促进大数据发展行动纲要的通知. [2015-09-05]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

Scientific Big Data—A Footstone of National Strategy for Big Data

GUO Huadong

(Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China)

Abstract Big data occupies the strategic high ground in the era of knowledge economies and also constitutes a new national and global strategic resource. It is a new pattern for scientific discovery with less dependence on causality and heavy dependence on data correlation. It has become a data-intensive scientific paradigm, following previous paradigms of empirical, theoretical and computational science. The paradigm has shifted the methodology of scientific research from theories and models based on causal analysis to comprehensive mechanistic scientific discovery including correlation analysis. As a branch of big data, scientific big data includes internal characteristics such as non-repeatability, high uncertainty, high dimensionality, and computational complexity. External characteristics include data type, data volume, data acquisition, and data analysis. All these characteristics bring new challenges for the techniques and methods of processing scientific big data. On the basis of the above analysis, we raise four recommendations: scientific cognition of scientific big data, construction of scientific big data infrastructure, establishment of a scientific data research center, and the structuring of a scientific big data academic platform.

Keywords big data, scientific big data, big earth data, data-intensive science



郭华东 中国科学院遥感与数字地球研究所研究员，博士生导师。中国科学院院士、俄罗斯科学院外籍院士、芬兰科学与人文院外籍院士、发展中国家科学院院士。主要从事遥感信息科学、雷达对地观测、数字地球等领域研究。现任国际数字地球学会主席、“联合国可持续发展目标技术促进机制10人组”成员、国际环境遥感委员会主席、联合国教科文组织国际自然与文化遗产空间技术中心主任、“数字丝路”国际科学计划主席、国家大数据专家委员会顾问、《国际数字地球学报》和《地球大数据》主编等职。现为中国科学院战略性先导科技专项（A类）“地球大数据科学工程”负责人。E-mail: hdguo@radi.ac.cn

GUO Huadong Professor of Institute of Remote Sensing and Digital Earth (RADI), Chinese Academy of Sciences (CAS), Academician of CAS, Foreign Member of the Russian Academy of Sciences, Foreign Member of the Finnish Society of Sciences and Letters, and Fellow of the World Academy of Sciences for the advancement of science in developing countries (TWAS). He presently serves as President of the International Society for Digital Earth (ISDE), Member of UN 10-Member Group to Support the Technology Facilitation Mechanism, Chairman of the International Committee on Remote Sensing of Environment (ICORSE), Director of the International Centre on Space Technologies for Natural and Cultural Heritage (HIST) under the Auspices of UNESCO, Chair of Science Committee of Digital Belt and Road Program (DBAR), Editor-in-Chief of the *International Journal of Digital Earth* and *Big Earth Data*. He served as President of ICSU Committee on Data for Science and Technology (CODATA). He specializes in remote sensing science and its applications, and has a series of achievements in remote sensing information mechanisms, radar for Earth observation, and Digital Earth science. He has published more than 600 papers and sixteen books, and is the principal awardee of sixteen domestic and international prizes. E-mail: hdguo@radi.ac.cn

■责任编辑：张帆